

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-110408

(43)Date of publication of application : 23.04.1999

(51)Int.Cl.

G06F 17/30

G06F 17/28

(21)Application number : 09-274323

(71)Applicant : SHARP CORP

(22)Date of filing : 07.10.1997

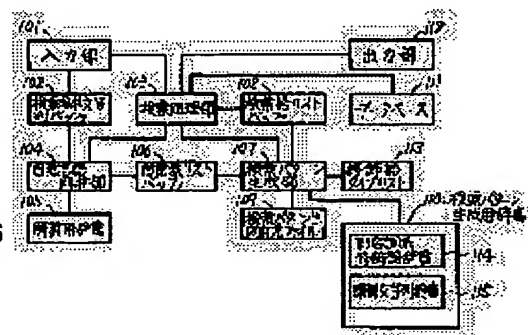
(72)Inventor : NAGAI TOSHIKAZU

(54) INFORMATION RETRIEVAL DEVICE AND METHOD THEREFOR

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an information retrieval device and method for generating retrieval words and phrases related to a retrieval requesting words and phrases to be the retrieval words and phrases, concerning the retrieval words and phrases inputted for retrieving related information in an information group.

SOLUTION: Retrieval request words and phrases inputted for retrieving information through an inputting part 101 are temporarily stored in a retrieval request character string buffer 102, decomposed into morphemic columns by a natural language analyzing part 104, and stored in a morpheme list buffer 106. A retrieval processing part 103 processes each character string on description of each morpheme in the buffer 106 by a retrieval pattern generating part 107 for newly generating retrieval words and phrases, and stores them in a retrieval word list buffer 108. Thus, a data base 111 can be retrieved based on each kind of retrieval words and phrases generated related with the inputted retrieval request words and phrases, and information pertinent to the retrieval request words and phrases can be retrieved, and outputted from an outputting part 112.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

Page Blank (uspto)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-110408

(43) 公開日 平成11年(1999) 4月23日

(51) Int.Cl.⁴

識別記号

F I

G 0 6 F 17/30
17/28

G 0 6 F 15/403 3 3 0 B
15/38 S
15/40 3 7 0 A
15/403 3 2 0 D
3 3 0 C

審査請求 未請求 請求項の数6 O L (全 13 頁)

(21) 出願番号 特願平9-274323

(22) 出願日 平成9年(1997)10月7日

(71) 出願人 000005049

シャープ株式会社

大阪府大阪市阿倍野区長池町22番22号

(72) 発明者 長家 利和

大阪府大阪市阿倍野区長池町22番22号 シ

ャープ株式会社内

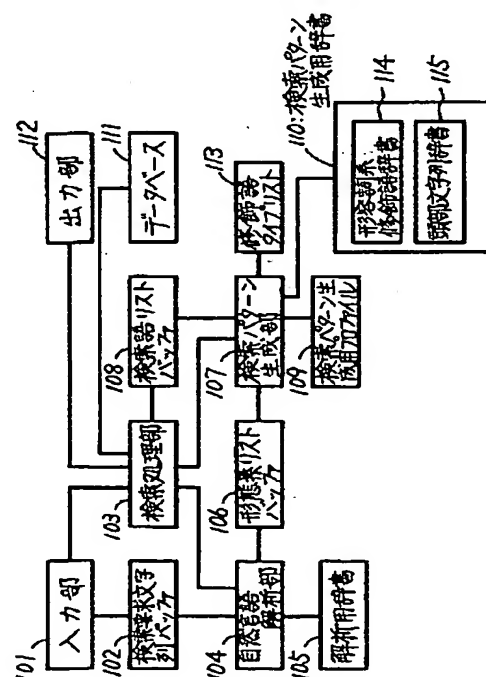
(74) 代理人 弁理士 深見 久郎

(54) 【発明の名称】 情報検索装置および方法

(57) 【要約】

【課題】 情報群中で関連する情報を検索するために入力される検索語句について、これと関連する語句を生成し検索語句とするような情報検索装置および方法を提供する。

【解決手段】 入力部101を介して情報を検索するために入力された検索要求語句は検索要求文字列バッファ102に一旦ストアされて、自然言語解析部104により形態素列に分解されて、形態素リストバッファ106にストアされる。検索処理部103はバッファ106中の各形態素の表記上の各文字列を検索パターン生成部107により処理して新たに検索語句を生成し検索語リストバッファ108にストアする。これにより、入力された検索要求語句に関連して生成される各種の検索語句によりデータベース111が検索されて検索要求語句に該当する情報が検索されて出力部112から出力される。



【特許請求の範囲】

【請求項1】 検索要求文字列を用いて、これに適合する情報をデータベース中から検索し提示する情報検索装置であって、

前記検索要求文字列を入力するための入力手段と、
入力された前記検索要求文字列を単語句である形態素の列に解析する形態素解析手段と、
前記形態素列における前記形態素のそれぞれについて、所定手順に従ってその文字列を操作して相互に結合し、新たな前記検索要求文字列として生成する生成手段とを備えた、情報検索装置。

【請求項2】 前記所定手順は、
体言の修飾語句となる前記形態素については、その頭部文字列を隣接する前記形態素の文字列に結合させる第1手順を含む、請求項1に記載の情報検索装置。

【請求項3】 前記頭部文字列は漢字からなることを特徴とする、請求項2に記載の情報検索装置。

【請求項4】 前記第1手順では、前記頭部文字列を前記隣接する形態素の文字列に結合させるときに、該結合が可能か否かに関する情報が参照されることを特徴とする、請求項2または3に記載の情報検索装置。

【請求項5】 前記所定手順は、
体言となる前記形態素については、その頭部文字列を隣接する前記形態素の文字列に結合させる第2手順を含む、請求項1～4のいずれかに記載の情報検索装置。

【請求項6】 検索要求文字列を用いて、これに適合する情報をデータベース中から検索し提示する情報検索方法であって、
前記検索要求文字列を入力するための入力ステップと、
入力された前記検索要求文字列を単語句である形態素の列に解析する形態素解析ステップと、
前記形態素列における前記形態素のそれぞれについて、所定手順に従ってその文字列を操作して相互に結合し、新たな前記検索要求文字列として生成する生成ステップとを備えた、情報検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は情報検索装置および方法に関し、特に、キーワードを入力してデータベースまたは文書情報を検索し、関連する情報を出力する情報検索装置および方法に関する。

【0002】

【従来の技術】従来の情報検索装置では、オペレータが入手したい情報が含まれであろうキーワードや任意の文字列を入力すると、それに適合した情報が提示される。入力するキーワードあるいは任意文字列は、それぞれ単独の語句を適宜に区切って入力する方法だけでなく、自然言語によって文として入力する方法をとるものもある。後者の場合にはそれを自然言語解析処理して検索語句に展開し、それに適合する情報が提示される。

【0003】どのような方法にしても検索処理の結果として提示される情報は、検索語句（前述のキーワードや任意文字列、あるいは自然言語解析処理して得られた検索語句）にマッチしたものであり、それらを含まないものは提示されない。

【0004】そこで、投入された検索語句に関して、それと関連のある関連語句に展開して、それをも自動的に検索語句として扱い処理することで、オペレータの意図をより満足する情報の提示が行なわれる。

【0005】この際、展開される関連語句は辞書情報として予め保存された情報から、検索語句に関連するものが採用される。

【0006】特開平5-342255号公報には、自然言語の形でデータベース内に保存されたすべての文を対象とする文書検索システムにおいて、検索語句と検索対象文書の双方を文単位で形態素解析処理し、さらに構文上の解析を行なって単語間の近接度を求めるとともに、加えて文中の各単語から展開可能な関連語句を収めた辞書情報を用いて単語を拡張し、それらを用いて検索することで検索漏れを防ぐ技術が開示される。

【0007】特開平7-210565号公報には、キーワードとして入力された検索語句を連想語辞書（1つの語に対して、それから連想される複数の語を対応させた辞書）を参照し関連語句に展開し、その内容をもってデータベースを検索する技術が開示される。

【0008】特開平8-137898号公報には、キーワードとして入力された語句がいかなる概念の語であるかを保存した概念辞書を参照し検索語句の概念と関連のある別概念の語句を得て、これによって検索語句を拡張し、文書データベースを検索する技術が開示される。

【0009】

【発明が解決しようとする課題】従来の技術における検索のためのキーワードとなる検索語句の拡張のためには辞書情報が必須となっている。また関連語句の辞書情報（それらは、各技術によって呼称が相違する。特開平5-342255号公報では類義語辞書、特開平7-210565号公報では連想語辞書、特開平8-137898号公報では概念辞書と呼ばれる）は、語単位について対応する語を複数個用意しなければ効果が得られず、辞書情報の容量が膨大なものになってしまう。

【0010】また、1つの語に1つの語を対応させる構成をとる（対応する語を複数個列挙することはない）ことから、複数の語（形態素）の組合せと1つの語の対応（たとえば「新しい車」→「新車」）、あるいは反対に1つの語から複数の語への対応（たとえば「新車」→「新しい車」）が表現できないので、そのような対応関係での検索語句の拡張ができない。

【0011】それゆえにこの発明の目的は、情報を検索するためのキーワードとして入力される検索要求文字列について、辞書を用いずこれと関連する文字列を複数個

生成し、これも情報検索のための検索要求文字列とする情報検索装置および方法を提供することである。

【0012】

【課題を解決するための手段】請求項1に記載の情報検索装置は、検索要求文字列を用いて、これに適合する情報をデータベース中から検索し提示する装置であって、検索要求文字列を入力するための入力手段と、入力された検索要求文字列を単語句である形態素の列に解析する形態素解析手段と、形態素列における形態素のそれぞれについて、所定手順に従ってその文字列を操作して相互に結合し、新たな検索要求文字列として生成する生成手段とを備えて構成される。

【0013】したがって、形態素解析手段により得られた各形態素の文字列を所定手順に従って操作し相互に結合することにより新たな検索要求文字列が生成されるので、関連語句に関する辞書情報を用いることなく検索要求語句の拡張を伴った情報検索が可能となって、辞書に関するコストを抑制することができる。また、生成手段による検索要求文字列の生成は語と語の1対1の変換ではなく文字列の操作および結合による多様な検索要求文字列の生成なので、検索者の意図をより満足する検索結果を得ることのできる検索要求文字列を準備することができる。

【0014】請求項2に記載の情報検索装置は、請求項1に記載の装置の生成手段における所定手順が、体言の修飾語句となる形態素については、その頭部文字列を隣接する形態素の文字列に結合させる第1手順を含んで構成される。

【0015】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞および連体詞などの体言修飾語について、その頭部の文字列を隣接する形態素の文字列に結合して新たな検索要求文字列を生成できるので、関連語句に関する辞書情報を用いることなく検索要求文字列の拡張を伴った検索が可能となって辞書情報に関するコストを抑制することができる。また、語と語の1対1の変換だけでなくより正確な検索要求文字列の拡張を伴った検索を可能にするので、検索者の意図をより満足する検索結果を得ることができる。

【0016】請求項3に記載の情報検索装置は、請求項2に記載の装置の所定手順における第1手順は、体言の修飾語句となる形態素については、その漢字からなる頭部文字列を隣接する形態素の文字列に結合させるよう構成される。

【0017】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞および連体詞などの体言修飾語について、その頭部の漢字部分を隣接する形態素の文字列に結合して新たな検索要求文字列を生成しこれを用いた検索を可能としている。それゆえ

に、関連語句に関する辞書情報を用いることなく検索要求文字列の拡張を可能にしているので、該装置における辞書に関するコストを抑制しながら多様な検索要求文字列を用いた検索が可能となる。また、検索要求文字列が拡張されることにより、検索者の意図をより満足する検索結果を得ることができる。

【0018】請求項4に記載の情報検索装置は請求項2または3に記載の装置における生成手段の所定手順の第1手順が、体言の修飾語句となる形態素の頭部文字列を隣接する形態素の文字列に結合させるときに、この結合が可能か否かに関する情報を参照するよう構成される。

【0019】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞および連体詞などの体言修飾語に関してその頭部文字列が隣接する形態素の文字列に結合が可能かどうかに関する情報を参照しながら結合して新たな検索要求文字列を生成し、これらを用いて検索を行なっているので、体言修飾語に関して、その頭部文字列を単純に隣接する形態素の文字列と結合させるよりも精度の高い検索要求文字列を生成することができ、検索者の意図をより満足する検索結果を得ることができる。

【0020】請求項5に記載の情報検索装置は、請求項1に記載の生成手段の所定手順は、体言となる形態素については、その頭部文字列を隣接する形態素の文字列に結合させる第2手順を含んで構成される。

【0021】したがって、生成手段において、特に漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている名詞句などの体言について、その頭部の文字を結合して新たな検索要求文字列を生成し検索が行なわれるので、関連語句に関する辞書情報を用いないで多様な検索要求文字列の拡張を伴った検索が可能になって、該装置における辞書情報に関するコストを抑制しながらより検索者の意図を満足する検索結果を得ることができる。

【0022】請求項6に記載の情報検索方法は検索要求文字列を用いて、これに適合する情報をデータベース中から検索し提示する方法であり、検索要求文字列を入力するための入力ステップと、入力された検索要求文字列を単語句である形態素の列に解析する形態素解析ステップと、形態素列における形態素のそれぞれについて、所定手順に従ってその文字列を操作して相互に結合し、新たな検索要求文字列として生成する生成ステップとを備えて構成される。

【0023】したがって、入力された検索要求文字列の各形態素の文字列を操作および結合することで、多様な検索要求文字列が新たに生成されるので、関連語句に関する辞書情報を用いることなく検索要求文字列の拡張を伴った検索が可能となって、辞書情報に関するコストを抑制しながら検索者の意図をより満足する検索結果を得

ることができる。

【0024】上述の情報検索方法は、生成ステップの所定手順が、体言の修飾語となる形態素については、その頭部文字列を隣接する形態素の文字列に結合させる第1手順を含んでもよい。

【0025】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞および連体詞などの体言修飾語について、その頭部の文字列を隣接する形態素の文字列に結合して新たな検索要求文字列を生成し検索を行なう際、関連語句に関する辞書情報を用いなくて検索要求文字列の拡張を伴った検索が可能となつて、辞書情報に関するコストが抑制される。また、正確な検索要求文字列の拡張を伴った検索が可能となつて、検索者の意図をより満足する検索結果を得ることができる。

【0026】また、前述の第1手順は、体言の修飾語からなる形態素については、その漢字からなる頭部文字列を隣接する形態素の文字列に結合させるように構成されてもよい。

【0027】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞ならびに連体詞などの体言修飾語の頭部の漢字部分を隣接する形態素の文字列に結合して新たな検索要求文字列を生成し検索を行なう際、関連語句に関する辞書情報を用いなくて検索要求文字列の拡張を伴った検索が可能となる。これにより、辞書情報に関するコストを抑制できる。さらに、多様な検索要求文字列の拡張を伴った検索が可能となつて検索者の意図をより満足した検索結果を得ることができる。

【0028】また、前述の第1手順は、体言修飾語となる形態素の頭部文字列を隣接する形態素の文字列に結合させるときに、結合が可能か否かに関する情報を参照するようにしてもよい。

【0029】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている形容詞、形容動詞および連体詞などの体言修飾語に関しては、その頭部文字列が隣接する形態素の文字列に結合可能か否かに関する情報が参照されながら、これらの結合が行なわれて新たな検索要求文字列が生成され検索が行なわれるので、体言修飾語に関して、その頭部文字列を単純に隣接する形態素の文字列と結合を行なうよりも精度の高い検索要求文字列の生成が可能となるので、検索者の意図をより満足する検索結果を得ることができる。

【0030】また、前述の生成ステップにおける所定手順は、体言となる形態素については、その頭部文字列を隣接する形態素の文字列に結合させる第2手順を含んでもよい。

【0031】したがって、漢字の持つ意味を基礎にした熟語の構成や語の結合による造語／省略語生成において重要な働きを持っている名詞句となる体言の形態素の頭部文字列を隣接する形態素の文字列と結合して新たな検索要求文字列を生成し検索を行なう場合に、関連語句に関する辞書情報を用いずに検索要求文字列の拡張を伴った検索が可能となる。これにより辞書情報に関するコストを抑制しながら多様な検索要求文字列の拡張が可能となる。また、これを用いて検索することにより、検索者の意図をより満足する検索結果を得ることができる。

【0032】

【発明の実施の形態】以下、この発明の実施の形態について図面を参照し説明する。この実施の形態では複数の異なる情報が格納されたデータベースから所望の情報の検索を要求する場合に、キーワードを示す文字列（以下、検索要求語句という）を入力すると、この検索要求語句に関連する複数の語句が生成されて（以下、検索語句という）、これも検索のためのキーワードに用いられ、データベースの検索がおこなわれる。このとき検索語句の生成は、後述するように漢字の持つ意味を基礎にした熟語の構成や語の結合による造語または省略語生成機能などに着目して、入力された検索要求語句の文字列を操作することにより行なわれる。

【0033】図1はこの発明の実施の形態による情報検索装置の機能構成図である。図において装置は外部から与えられる検索要求語句を入力するための入力部101、入力された検索要求語句を格納する検索要求文字列バッファ102、入力された検索要求語句をもとにデータベースを検索してその結果を出力するための検索処理部103、入力された自然言語による検索要求語句を解析するための自然言語解析部104、自然言語解析部104が自然言語を解析する際に参照する情報を格納した解析用辞書105、自然言語解析部104が解析結果として生成する形態素情報を格納するための形態素リストバッファ106、自然言語解析部104の処理結果に基づき検索語句を生成するための検索パターン生成部107、検索パターン生成部107が生成する検索語句を格納するための検索語リストバッファ108、検索パターン生成部107が検索語句を生成する際に参照する所定情報が格納された検索パターン生成用プロフィール109、検索パターン生成部107が検索語句を生成する際に参照する辞書情報を格納した検索パターン生成用辞書110、検索対象となる情報を格納したデータベース111、検索処理部103がデータベース111中を検索して得られた情報を出力するための出力部112、および修飾語タイプリスト113を含む。検索パターン生成用辞書110は形容詞系修飾語辞書114および頭部文字列辞書115を含む。なお、修飾語タイプリスト113、形容詞系修飾語辞書114および頭部文字列辞書115の詳細は後述する。

【0034】図2は図1の形態素リストバッファ106の構成例の図である。図2では形態素リストバッファ106に格納される自然言語解析部104の解析結果である形態素情報の形式が示される。形態素リストバッファ106では各形態素の情報がレコードR_i (i=1、2、3、…)に格納される。レコードR_iのそれぞれはフィールドF₁～F₃を含み、フィールドF₁には形態素の品詞情報I₁、フィールドF₂には位置情報I₂およびフィールドF₃には長さ情報I₃が格納される。品詞情報I₁は対応する形態素の品詞(名詞、形容詞など)を示す情報であり、位置情報I₂は対応する形態素が検索要求文字列バッファ102中で存在する箇所を特定するためのバッファ102先頭からのバイトオフセット情報である。長さ情報I₃は対応する形態素がバッファ102内で占有するバイト長情報である。

【0035】図3は、図1の検索語リストバッファ108の構成例の図である。図3では検索語リストバッファ108に格納される検索パターン生成部107の処理結果である検索語句の情報の形式が示される。検索語リストバッファ108には検索語句のそれぞれについて不定長のレコードR_i (i=1、2、3、…)が設けられる。レコードR_iのそれぞれはフィールドF₁およびF₂を含む。フィールドF₁には長さ情報I₄が格納されて、情報I₄は対応する検索語句の長さ、即ち対応するフィールドF₂の文字列情報I₅の長さを示す。文字列情報I₅は対応する検索語句の実際の文字列を示す。

【0036】図4は図1の検索パターン生成用プロフィール109の構成例の図である。図4の検索パターン生成用プロフィール109には、形容詞系修飾語の頭部処理の方法を示す設定情報と名詞系修飾語の頭部処理の方法を示す設定情報が格納される。ここで「形容詞系修飾語」とは形容詞、形容動詞および連体詞などの体言修飾機能を有する自立語と定義する。また「名詞系修飾語」とは単独の名詞または助詞「の」を伴った名詞であると定義する。「頭部」とはそれぞれの語の先頭から始まる部分文字列であると定義する。

【0037】形容詞系修飾語の頭部処理の方法には、「無条件に処理」、「形容詞系修飾語頭部辞書を参照して処理」、「頭部文字列辞書を参照して処理」および「漢字部分のみ処理」の方法があるものとし、それぞれを値0、1、2、3で表現して図4の形容詞系修飾語の頭部処理の方法としていずれかの値を設定情報として格納する。図4の例では形容詞系修飾語の頭部処理の方法が無条件に処理に設定されている。また、名詞系修飾語の頭部処理の方法には、「すべての字種を処理」、「漢字のみを処理」の各方法があるものとして、それぞれを値0、1で表現して図4の名詞系修飾語の頭部処理の方法としていずれかの値を設定情報として格納する。図4の例では名詞系修飾語の頭部処理の方法はすべての字種を処理に設定されている。

【0038】図5は、図1の検索パターン生成部107のブロック図である。検索パターン生成部107は検索パターンである検索語句を生成する処理を実行する処理部501および処理部501がその処理中に再帰的処理を行なう際に利用するワーク領域である再帰テーブル502を含む。

【0039】図6は図5の再帰テーブル502の構成例の図である。図において再帰テーブル502は再帰テーブルポインタt_{abp}で指示されるレコードR_iを含む。レコードR_iのそれぞれは図2に示された形態素リストバッファ106中の各形態素に対応して設けられる。再帰テーブル502の各レコードには対応する形態素の形態素リストバッファ106中でのレコード位置を示すレコード位置情報pos、対応する形態素の長さを示す形態素長さ情報len、再帰処理中での対応する形態素についての現在処理中の頭部の長さを示す現在処理中の頭部長さ情報curおよび対応する形態素の修飾語タイプ(本実施の形態の場合、形容詞系修飾語および名詞系修飾語のいずれかのタイプ)を示す形態素の修飾語タイプ情報typeを含む。

【0040】次に、オペレータにより入力部101から検索要求語句として「新しい税制の研究会」という文字列が入力されたとして、一連の処理の流れを説明する。

【0041】図7は、図1の検索処理部103の処理のフローチャートである。図8は、図1の検索要求文字列バッファ102の構成例の図である。図9は、図1の形態素リストバッファ106の内容の一例を示す図である。

【0042】図7において、まず、検索要求語句の入力を入力部101に対して要求し、その入力を使って検索要求語句を獲得する(S701)。今の場合、「新しい税制の研究会」という文字列が検索要求文字列バッファ102に格納される。図8にその状態が示される。

【0043】次に、入力された検索要求語句の文字列に対して自然言語解析処理を行なうように自然言語解析部104に処理が要求される(S702)。今の場合は、自然言語解析処理として形態素解析処理が行なわれるとする。形態素解析処理は一般に行なわれる方法が採用されるので説明は省略する。

【0044】自然言語解析部104の処理の結果は形態素リストバッファ106に図2に示される形式で格納される。今の場合の形態素リストバッファ106の内容が図9に示される。図9では形態素レコードポインタ1rpで指示される形態素「新しい」、「税制」、「の」、「研究」、「会」に分割、認識されて先に説明した内容に従って各形態素に対応して情報I₁～I₃が格納される。図9中の位置情報I₂の値が図8での各形態素の開始位置に相当する。なお、図9中の右端の()の内容は対応する形態素がどれであるかをわかりやすく示すために便宜的に付与したものである。

【0045】自然言語解析部104によって検索要求語句が形態素に分割、認識された後、それをもとに検索語句の生成を行なうように検索パターン生成部107に処理が要求される(S703)。その結果は検索語リストバッファ108の図3に示した形式で検索語句のリストとして生成されるので、それをもとにデータベース111に対して検索を要求し(S704)、その結果を出力部112に出力する(S705)。なお、データベース111への検索の要求および出力部112への出力方法は一般的な方法であればよいので説明は省略する。

【0046】図10は図1の検索パターン生成部107の処理のフローチャートである。検索パターン生成部107の処理前の状態として形態素リストバッファ106に図9のようなデータが設定されている。まず、処理対象となる形態素リストバッファ106のレコードRi(形態素)を示す形態素レコードポインタlrpを形態素リストバッファ106の先頭レコードを指すように初期化する(S1001)。

【0047】次にポインタlrpの指すレコードRiから対応する形態素は付属語であるかどうかを品詞情報I1から判断する(S1002)。付属語である場合はポインタlrpを次のレコードRiを指すように更新して、次の形態素の処理を行なう準備をし(S1004)、付属語でない場合は後続の処理(S1003)を行なう。今の場合、ポインタlrpの指すレコードRiの形態素「新しい」は付属語でないので後続の処理(S1003)が行なわれる。

【0048】形態素が付属語でない場合には検索語句を生成する処理が行なわれ(S1003)、その結果は検索語リストバッファ108に格納される。その後、形態素が最後のものであるかどうかをチェックして(S1005)、次の形態素があれば処理をS1002に戻してループする。形態素が最後のものであれば処理は終了する。

【0049】図11は図10の検索語生成処理のフローチャートである。検索語生成処理開始直前の状態では、形態素レコードポインタlrpが処理すべき形態素に対応する形態素リストバッファ106中のレコードRiを指している。

【0050】まず、再帰テーブル502の初期セットが行なわれる(S1101)。再帰テーブル502の初期セットが終了すると、図6において再帰テーブル502の対応するレコードRiのレコード位置情報pos、形態素長さ情報lenおよび形態素の修飾語タイプ情報typeの値が形態素リストバッファ106の対応するレコードRiの内容に基づいて設定される。再帰テーブル502のレコード位置情報posには対応する形態素のレコードRiが形態素リストバッファ106(図9参照)のどのレコードRiに対応するかを示し、情報lenは対応する形態素の長さ情報I3を示し、情報type

eは対応する形態素の修飾語タイプ(今の場合、形容詞系修飾語は1、名詞系修飾語は2で表現するものとする)を示す。再帰テーブル502の最後のレコードRiのレコード位置情報posの値は-1であり、テーブル502の終端を示す。

【0051】再帰テーブル502の初期セットが終了すると、次に再帰テーブル502のレコードRiを指す再帰テーブルポインタtabpがテーブル502の先頭レコードを指すように初期化される(S1102)。ここから、再帰テーブル502中の各形態素について繰返し検索語の生成処理が行なわれる。

【0052】まず、テーブルポインタtabpの指すレコードRiの現在処理中の頭部長さ情報curを0で初期化し(S1103)、次にテーブルポインタtabpの指すレコードRiの情報typeによってその修飾語タイプをチェックする(S1104)。その修飾語タイプによって、名詞系修飾語であれば名詞系修飾語処理を、形容詞系修飾語であれば形容詞系修飾語処理が行なわれる(S1105、S1106)。これらの処理を終えた段階では、処理対象の形態素の頭部の処理のための文字列として選択された文字数がテーブルポインタtabpによって示されるレコードRiの現在処理中の頭部長さ情報curにセットされる。情報curの値が-1であるときには選択すべき頭部の処理がなかったことを示す。各処理が終了するとテーブルポインタtabpの示すレコードの現在処理中の頭部長さ情報curの値によって頭部文字列が選択されたかどうかチェックされる(S1107)。

【0053】処理対象となっている形態素について頭部文字列が選択されたならば、テーブルポインタtabpの指す次のレコードRiを参照して、そのレコードRiが再帰テーブル502の最終レコードであるかどうかによって次の形態素が存在するかどうか確認する(S1108)。次の形態素が存在するならば、テーブルポインタtabpを1つ進め(S1109)、次の形態素を処理する準備を行なって処理S1103に戻る。

【0054】次の形態素が存在しないのならば、1つの検索語句の生成が終了したので、検索語リストバッファ108に生成された検索語句をセットして(S1110)、セットできれば処理S1104に戻り、セットできなければエラーとして処理を終了するため処理S1114に進む(S1111)。

【0055】処理S1107において頭部文字列の選択処理が終了したという状態が検知されると、テーブルポインタtabpを1つ前のレコードRiに戻し(S1112)、処理S1104に戻る。

【0056】テーブルポインタtabpが既に再帰テーブル502の先頭のレコードを指しており戻せないときは処理全体を終了するため、形態素レコードポインタlrpを検索語生成処理で処理対象にしていた形態素列が

最後の形態素に対応するレコードRiを指すようにセットし(S1114)、処理を終了する。

【0057】図12は図11の再帰テーブル初期セット処理のフローチャートである。図13は図1の修飾語タイプリストの構成例の図である。

【0058】図12の再帰テーブル初期セット処理においては処理のため再帰テーブルポインタtabpが初期化され(S1201)、作業用形態素リストポインタpが初期化される(S1202)。次に、ポインタpの指す形態素の品詞を修飾語タイプリスト113中から検索し、そこに示された修飾語タイプを取得する(S1203)。修飾語タイプリスト113は図13に示されるように品詞と修飾語タイプの対応表である。

【0059】得られた修飾語タイプに従って処理を分岐し(S1204)、名詞系修飾、形容詞系修飾の場合はそれぞれの値をテーブルポインタtabpの指すレコードRiの形態素の修飾語タイプ情報typeにセットする(S1205、S1206)。名詞系修飾および形容詞系修飾のいずれでもない場合には次の形態素について処理するため処理S1209に移行する。

【0060】対応する形態素が名詞系修飾および形容詞系修飾のいずれかの修飾語であった場合には、テーブルポインタtabpが指すレコードRiのレコード位置情報posおよび形態素長さ情報lenのセットが行なわれ(S1207、S1208)、次の形態素の処理のためポインタl rpおよびtabpを1つ進め(S1209)、形態素の存在を確認すれば(S1210)、処理S1203に戻る。

【0061】最後の形態素が存在しないときは、端末を示す値-1をテーブルポインタtabpにセットして(S1211)、処理を終了する。

【0062】図14は図11の名詞系修飾語処理のフローチャートである。図14の名詞系修飾語処理の開始時点での再帰テーブルポインタtabpが指すレコードRiの頭部長さ情報curの値を1だけ加算し(S1401)、その値と対応する形態素長さ情報lenの値を比較する(S1402)。長さ情報curが長さ情報lenを超えていれば処理を終了するため処理S1405へ移行する。

【0063】長さ情報curが長さ情報lenを超えていなければ、次に検索パターン生成用プロファイル109の内容から名詞系修飾語の頭部処理の方法を確認し処理を分岐させる(S1403)。

【0064】確認された頭部処理の方法が「すべての字種を処理」の場合にはそのまま処理を終了するが、「漢字のみを処理」の場合には形態素の対応する頭部長さ情報cur-1文字目(文字列中の文字の位置は0から数えとすると)が漢字か否かをチェックする(S1404)。漢字であれば頭部として選択する意味でそのまま処理を終了するが、漢字でなければこれ以上頭部を選択

できないことを意味するため現在処理中の頭部長さ情報curに-1をセットして(S1405)、処理を終了する。

【0065】図15は図11の形容詞系修飾語処理のフローチャートである。図16は、図1の形容詞系修飾語辞書の構成例の図である。図15において形容詞系修飾語処理の開始時点での再帰テーブルポインタtabpが指すレコードRiの現在処理中の頭部長さ情報curの値を1だけ加算し(S1501)、その値と対応する形態素長さ情報lenの値を比較する(S1502)、長さ情報curが長さ情報lenを超えていれば処理を終了するため処理S1506へ移行する。

【0066】長さを超えていなければ、次に、検索パターン生成用プロファイル109の内容から形容詞系修飾語の頭部処理の方法を確認し処理を分岐させる(S1503)。この場合「無条件に処理」の場合はそのまま処理を終了するが「形容詞系修飾語辞書を参照して処理」と「頭部文字列辞書を参照して処理」と「漢字部分のみ処理」の場合は次のように処理される。

【0067】まず、「形容詞系修飾語辞書を参照して処理」の場合、処理中の形態素が形容詞系修飾語辞書114に存在するかどうか確認する(S1504)。形容詞系修飾語辞書は図16に示されるような形態素Bと頭部文字列として有効な最大長さLの対応を保持した辞書情報である。処理中の形態素が辞書114に存在しない場合は、頭部選択処理の終了を意味するため再帰テーブルポインタtabpの指すレコードRiの現在処理中の頭部長さ情報curの値に-1をセットして(S1506)、処理を終了する。

【0068】「漢字部分のみ処理」の場合、現在処理対象としている形態素の頭部の漢字が接続する長さを求めてそれを有効頭部最大長として保持し(S1508)、有効頭部最大長と現在の頭部中の比較処理を行ない(S1505)、超えていなければそのまま処理を終了し、超えていれば頭部選択処理の終了を意味するため再帰テーブルポインタtabpの指すレコードRiの現在処理中の頭部長さ情報curに-1をセットして(S1506)、処理を終了する。

【0069】図17は図11の検索語リストセット処理のフローチャートである。図18は図6の再帰テーブル502の内容の一例を示す図である。図19は図3の検索語リストバッファ108の内容の一例を示す図である。図20は図3の検索語リストバッファ108の内容のその他の例を示す図である。図21は図3の検索語リストバッファ108の内容のさらなるその他の例を示す図である。

【0070】図17の検索語リストセット処理において、まず、作業用の再帰テーブルポインタpを再帰テーブル502の先頭レコードを指すようにセットする(S1701)。次にポインタpが指すレコードRiがテ

ブル502の最終のレコードであるかどうかをチェックし(S1702)、最終レコードなら処理を終了するが、最終レコードでなければ、ポインタpの指すレコードRiの現在処理中の頭部長さ情報curを利用して形態素の先頭から情報cur-1文字目までの文字列を検索語リストバッファ108に出力する(S1703)。検索語リストバッファ108への出力のは図3に示された構成で検索語句を追加していく単純な処理なので説明は省略する。出力後、ポインタpを進めて処理S1702に戻り、すべての形態素について処理を繰返す。

【0071】上述した検索語生成処理がどのように働くか、「新しい税制の研究会」が投入された場合で解説する。なお、検索パターン生成用プロファイル109の内容は図4に示された状態、すなわち名詞系修飾語の頭部処理の方法は「すべての字種を処理」および形容詞系修飾語の頭部処理の方法は「無条件に処理」が設定されているものとする。

【0072】検索語生成処理開始直前の状態では、形態素レコードポインタlrpが処理すべき形態素に対応する形態素リストバッファ106中のレコードRiを指し示している。今の場合、図2の最初のレコードR0、すなわち「新しい」に対応するレコードR0を指している。

【0073】まず、再帰テーブル502の初期セットが行なわれ(S1101)、図9の内容から図18に示されたように設定される。

【0074】なお、図18において右端の()内は対応する形態素がどれであるかをわかりやすく示すために便宜的に付与されたものである。処理S1102およびS1103と進み、処理S1104において対応の形態素の修飾語タイプがチェックされる。今の場合形態素「新しい」の品詞情報I1は形容詞を示すので、再帰テーブル初期セット処理中で参照される修飾語タイプリスト113から得られる情報として、(図18にあるように)その修飾語タイプは「形容詞系修飾語」と設定されているので、形容詞系修飾語処理が行なわれる(S1106)。検索パターン生成用プロファイル109の形容詞系修飾語の頭部処理の方法が「無条件に処理」に設定されているので、形容詞系修飾語処理において再帰テーブル502の現在処理中の頭部長さ情報curの値が1を加えた値になって戻ってくる(S1501)。その後、処理はS1107、S1108およびS1109と進みテーブルポインタtabpが次の形態素のレコードRiを指すようになり、処理S1104に戻る。

【0075】次に示される新しい形態素「税制」は名詞系修飾語であることが同様にその品詞情報I1と修飾語タイプリスト113から判明し、名詞系修飾語処理が行なわれる(S1105)。

【0076】検索パターン生成用プロファイル109の名詞系修飾語の頭部処理の方法が「すべての字種を処

理」に設定されているので、名詞系修飾語処理において再帰テーブル502の対応する長さ情報curの値が1を加えられた値になって戻ってくる(S1401)。先ほどと同様に処理がS1107、S1108およびS1109と進み、テーブルポインタtabpが次の形態素のレコードRiを指し示すようになり、処理S1104に戻る。

【0077】次の新しい形態素「研究」も名詞系修飾語であることが同様にその品詞情報I1と修飾語タイプリスト113から判明し、名詞系修飾語処理が行なわれる(S1105)。

【0078】検索パターン生成用プロファイル109の名詞系修飾語の頭部処理の方法が「すべての字種を処理」に設定されているので、名詞系修飾語処理において再帰テーブル502の対応するレコードR2の現在処理中の頭部長さ情報curの値が1を加えた値になって戻ってくる(S1401)。先ほどと同様に処理がS1107およびS1108と進む。このとき、次の形態素がないことがわかるので(次のレコードR3のレコード位置情報posにテーブル502の終端を意味する-1が格納されているので)、処理S1110に進んで、検索語リストセット処理が行なわれる。

【0079】再帰テーブル502の各形態素(各レコードR0~R2のそれぞれ)の現在処理中の頭部長さ情報curの値は、それぞれ「1」、「1」、「1」となっており、ここから検索語である文字列「新税研」を得ることができる。そのときの検索語リストバッファ108の内容が図19に示される。

【0080】その後、処理S1111からS1104に戻るが、この時点でのテーブルポインタtabpは再帰テーブル502の最初に登録された形態素「研究」のレコードR2を指したままであり、対応の現在処理中の頭部長さ情報curの値も「1」のままなので、処理S1105において長さ情報curが「2」になって戻ってくる。以降は同様に処理S1107およびS1108と進む、やはり最後に登録された形態素であるので検索語句が出力されるが、今回は再帰テーブル502の各形態素の各レコードRiの頭部長さ情報curの値は、それぞれ「1」、「1」、「2」となっており、ここから検索語句として文字列「新税研究」を得ることができる。そのときの検索語リストバッファ108の状態が図20に示される。

【0081】さらにもう一度処理を繰返すと、今回は形態素「研究」の形態素長さ情報lenは2を示すことから、処理S1401で加算された現在処理中の頭部長さ情報curの値(=3)が長さ情報len(=2)の値を超えてしまい、頭部選択処理の終了を意味するように現在処理中の頭部長さ情報curに値-1がセットされて処理S1107に戻る。したがって、検索語は出力されず処理S1112においてテーブルポインタtabp

が1つ前のレコードR1を指すように戻され、処理対象の形態素が「税制」となって、同様にして処理S1104から処理が開始される。

【0082】この時点では、テーブルポインタt a b pが指すレコードR1の現在処理中の頭部長さ情報c u rの値は「1」のままで、これが処理S1105において加算されて値「2」になることから、形態素「税制」から頭部の2文字が選択されることになり、先ほどと同様に形態素「研究」について繰返し処理をすることにより、検索語として文字列「新税制研」および「新税制研究」を得ることができる。

【0083】この時点で処理S1107で同様に対象とする形態素が1つ戻り、形態素「新しい」のレコードR0を指すようにテーブルポインタt a b pがセットされる。その後は、同様に後続の形態素について処理が続き、最後に「新しい」について処理が続けられなくなり、加えてテーブルポインタt a b pが前に戻せなくなるまで処理が続いて、最終的に図21に示されるような検索語生成結果を得ることができる。

【0084】以上述べてきたように、入力された図8の検索要求語句から図21に示された多くの検索語句を生成することができる。

【0085】次に、図4に示された検索パターン生成用プロファイル109の形容詞系修飾語の頭部処理の方法が「無条件に処理」でなく、「形容詞系修飾語辞書を参照して処理」に設定されている場合について説明する。

【0086】この場合、前述の処理の流れとほとんど変わるところはないが、図15に示された形容詞系修飾語処理の中で、形容詞系修飾語辞書114を参照するようになる(S1504)。形容詞系修飾語辞書114は図16に示されたような内容であったとすると、形態素「新しい」では有効頭部最大長が「1」となっていることから、対応の現在処理中の頭部長さ情報c u rの値が1以外のとき(たとえば2または3)は、検索語の頭部処理が終了したものとする情報が与えられるので、図21中の検索語「新し税研」または「新しい税研」は出力されない。

【0087】同様に、図4に示された検索パターン生成用プロファイル109の形容詞系修飾語の頭部処理の方法が「無条件に処理」でなく、「頭部文字列辞書を参照して処理」に設定されている場合、前述の処理の流れとほとんど変わるところはないが、図15に示された形容詞系修飾語処理の中で、頭部文字列辞書115を参照するようになる(S1507)。頭部文字列辞書115は前述のように頭部として有効な文字列を列挙した単純な辞書であるので、たとえば「新」が存在し、「新し」や「新しい」が頭部文字列辞書115に存在しなければ、図21中の検索語「新し税研」または「新しい税研」が出力されない。

【0088】同様に、図4に示された検索パターン生成

用プロファイル109の形容詞系修飾語の頭部処理の方法が、「無条件に処理」でなく、「漢字部分のみ処理」に設定されている場合、前述の処理の流れとほとんど変わるところはないが、図15に示された形容詞系修飾語処理の中で、形態素の頭部の漢字連接数を有効頭部長として処理する(S1505)。これにより、たとえば形態素「新しい」については「新」は有効であるが、「新し」や「新しい」は有効でないので、図21中の検索語「新し税研」または「新しい税研」は出力されない。

【0089】上述したように、本実施の形態において検索要求語句から漢字の持つ意味を基礎にした熟語の構成や語の結合による造語または省略語生成機能などにより、検索要求語句の文字列が操作された多様な検索語句が生成される。それは、たとえばこれ以外にも、「日本弁護士連合会」→「日弁連」、「特別な委員会」→「特別委員会」、「大蔵省の原案」→「大蔵原案」など、略語や省略語、または「新しい車」→「新車」などの漢字の意味を基礎にした熟語などを生成することができる。

【0090】なお、本実施の形態で通常の類義辞書や概念辞書を併用してさらに検索語句を拡張するようにしてもよい。

【0091】図22は、この発明の実施の形態による情報検索装置のハードウェア構成図である。

【0092】図22において情報検索装置は装置全体の制御を司るCPU(中央処理装置)2201、装置に検索要求語句を含む各種指示を入力するためのキーボード2202、プログラムや静的データを格納する不揮発性メモリであるROM2203、演算データや生成データを格納する揮発性メモリであるRAM2204、ハードディスクなどの外部記憶装置2205、表示のための情報を展開するVRAM2206、および情報を表示するためのCRT2207を含む。

【0093】CPU2201は図1の検索処理部103、自然言語解析部104および検索パターン生成部107として機能する。キーボード2202は図1の入力部101として機能する。ROM2203は解析用辞書105、検索パターン生成用プロファイル109、検索パターン生成用辞書110として機能する。RAM2204は検索要求文字列バッファ102、形態素リストバッファ106および検索語リストバッファ108として機能する。外部記憶装置2205はデータベース111として機能する。外部記憶装置2205は場合によっては、解析用辞書105、検索パターン生成用プロファイル109および検索パターン生成用辞書110として機能してもよい。VRAM2206およびCRT2207は図1の出力部112として機能する。

【図面の簡単な説明】

【図1】この発明の実施の形態による情報検索装置の機能構成図である。

【図2】図1の形態素リストバッファの構成例の図であ

る。

【図3】図1の検索語リストバッファの構成例の図である。

【図4】図1の検索パターン生成用プロファイルの構成例の図である。

【図5】図1の検索パターン生成部のブロック図である。

【図6】図5の再帰テーブルの構成例の図である。

【図7】図1の検索処理部の処理のフローチャートである。

【図8】図1の検索要求文字列バッファの構成例の図である。

【図9】図1の形態素リストバッファの内容の一例を示す図である。

【図10】図1の検索パターン生成部の処理のフローチャートである。

【図11】図10の検索語生成処理のフローチャートである。

【図12】図11の再帰テーブル初期セット処理のフローチャートである。

【図13】図1の修飾語タイプリストの構成例の図である。

【図14】図11の名詞系修飾語処理のフローチャートである。

【図15】図11の形容詞系修飾語処理のフローチャートである。

【図16】図1の形容詞系修飾語辞書の構成例の図であ

る。

【図17】図11の検索語リストセット処理のフローチャートである。

【図18】図6の再帰テーブルの内容の一例を示す図である。

【図19】図3の検索語リストバッファの内容の一例を示す図である。

【図20】図3の検索語リストバッファの内容のその他の例を示す図である。

【図21】図3の検索語リストバッファの内容のさらなるその他の例を示す図である。

【図22】この発明の実施の形態による情報検索装置のハードウェア構成図である。

【符号の説明】

101 入力部

102 検索要求文字列バッファ

103 検索処理部

104 自然言語解析部

106 形態素リストバッファ

107 検索パターン生成部

108 検索語リストバッファ

109 検索パターン生成用プロファイル

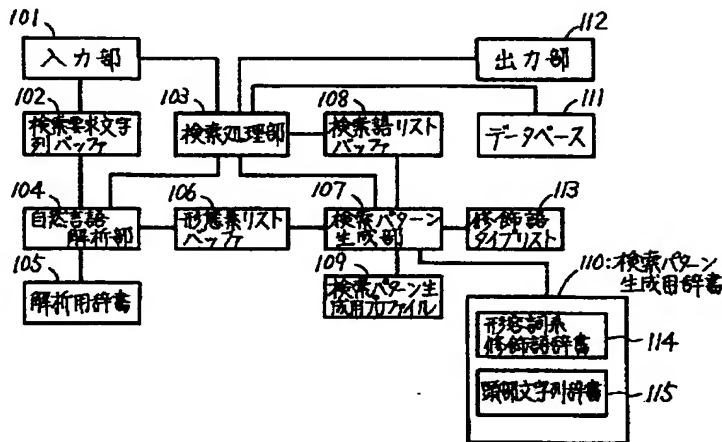
110 検索パターン生成用辞書

112 出力部

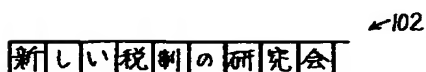
113 修飾語タイプリスト

なお、各図中同一符号は同一または相当部分を示す。

【図1】



【図8】



【図2】

F1	F2	F3	
I1	I2	I3	R0
			R1
			R2
			R3
⋮	⋮	⋮	⋮
			Ri
⋮	⋮	⋮	⋮

F1, F2, F3: フィールド

Ri (i=0, 1, 2, ...): レコード

I1: 品詞情報

I2: 位置情報

I3: 長さ情報

【図3】

F1	F2	
I4	I5	r0
		r1
		r2
⋮	⋮	⋮
		ri
⋮	⋮	⋮

$ri(i=0,1,2,\dots)$:レコード

I4:長さ情報

I5:文字列情報

【図6】

	F1	F2	F3	F4	
tabp	pas	len	cur	type	R0
					R1
					R2
					⋮

pos:ワード位置情報
len:形態素長さ情報
cur:現在処理中の頭部長さ情報
type:形態素の修飾語タイプ情報
tabp:再帰テールポインタ

【図9】

I1	I2	I3	
修飾	0	3	(新しい)
名詞	3	2	(税制)
助詞	5	1	(の)
名詞	6	2	(研究)
助詞	8	1	(会)

lrp:形態素ロードポインタ

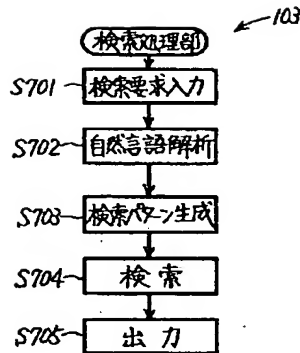
【図18】

pos	len	cur	type	
0	3	(不定)	1	(新しい)
1	2	(不定)	2	(税制)
3	2	(不定)	2	(研究)
-1				

【図4】

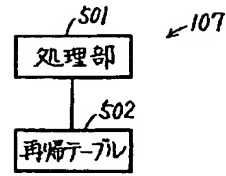
形容詞系修飾語の頭部処理方法	0
名詞系修飾語の頭部処理方法	0

【図7】



【図10】

【図5】



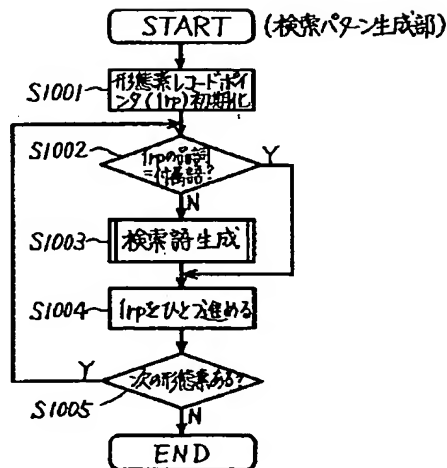
【図13】

H	T
品詞	修飾語タイプ
名詞	2
形容詞	1
形動	1
連体詞	1

【図16】

B	L
形態素	有効領域最大長
⋮	⋮
新しい	1
特別	2
⋮	⋮
⋮	⋮

【図19】

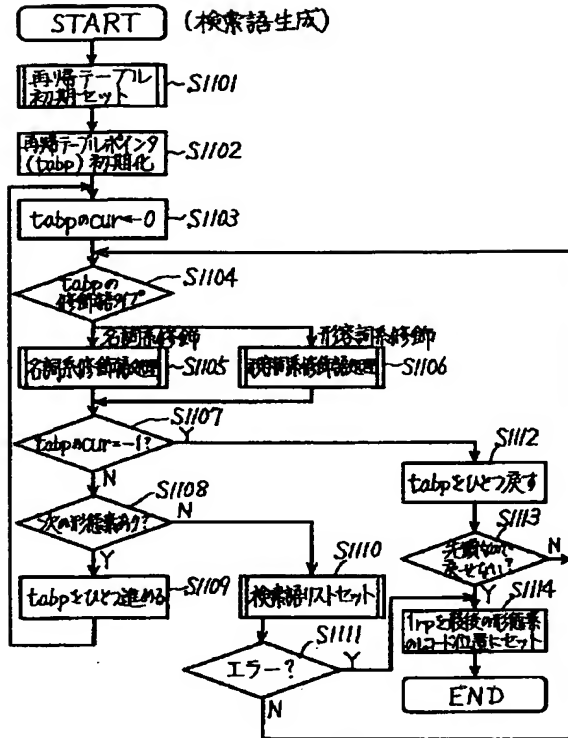


長さ	文字列
3	新税研

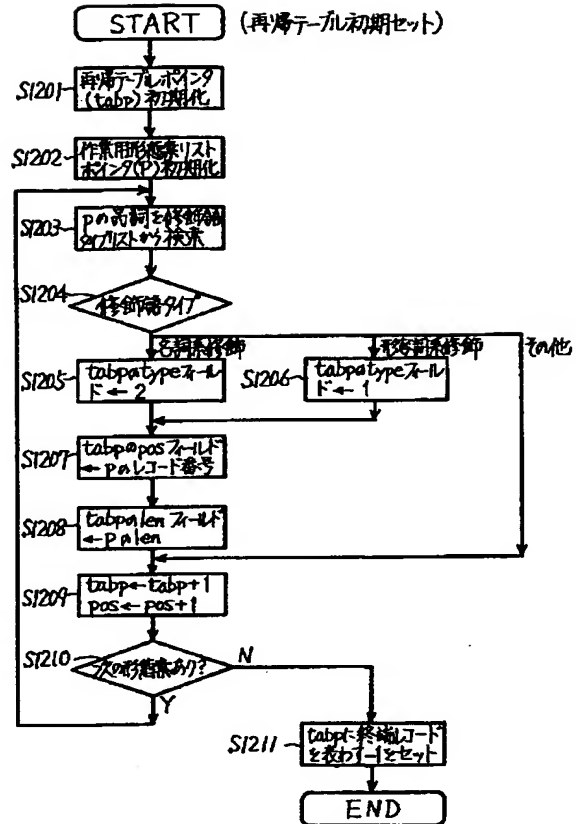
【図20】

長さ	文字列
3	新税研
4	新税研究

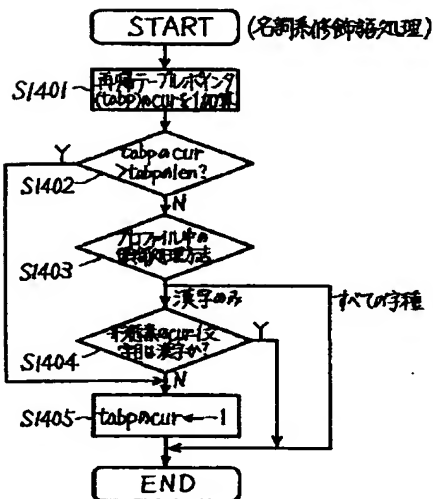
【図11】



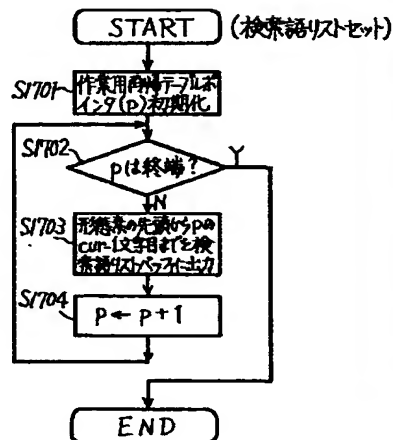
【図12】



【図14】



【図17】

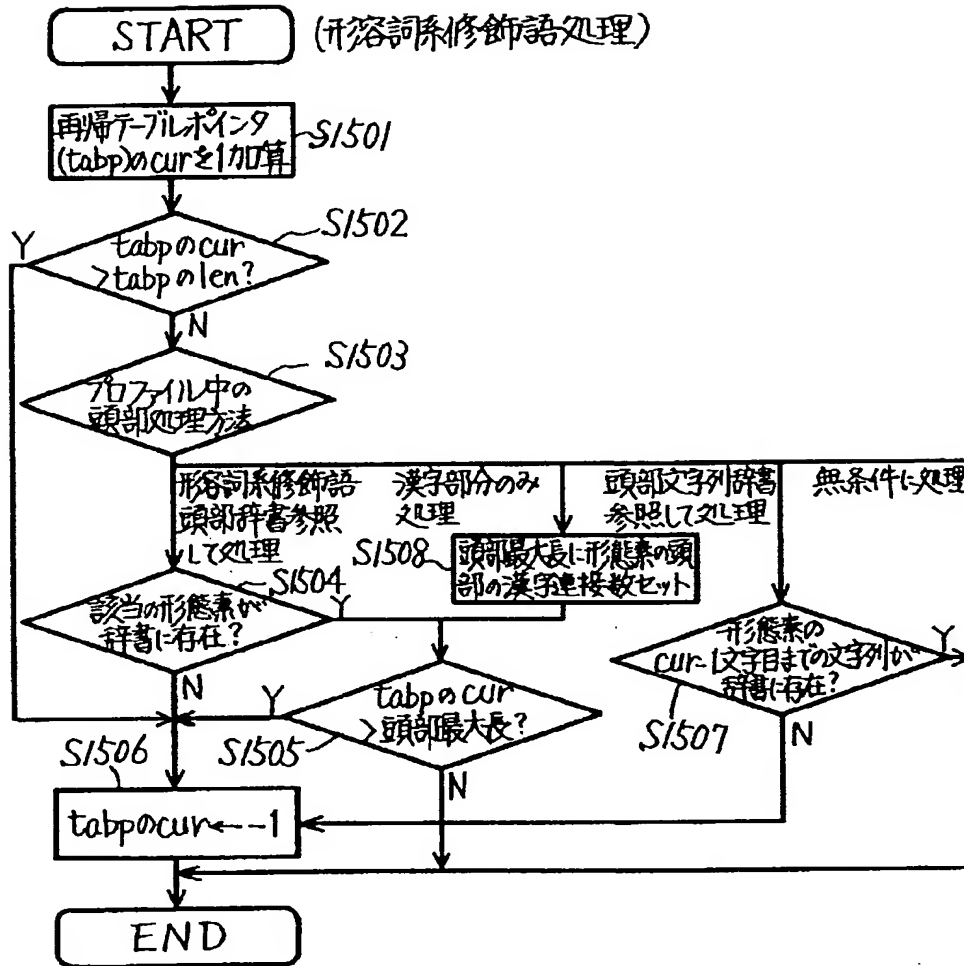


【図21】

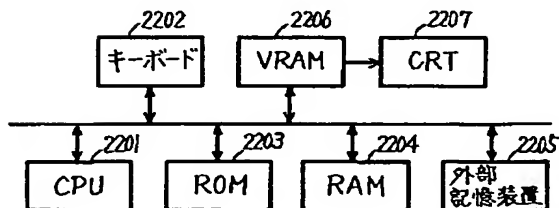
表で 文字列 ←108

3	新税研
4	新税研究
4	新税制研
4	新税制研究
3	新し税研
4	新し税研
4	新し税制研
4	新し税制研
3	新しい税研
4	新しい税研究
4	新しい税制研
4	新しい税制研究

【図15】



【図22】



This Page Blank (uspto)